Predicting phoneme and word recognition in noise using a computational model of the auditory periphery

Arturo Moncada-Torres[1], Astrid van Wieringen[1], Ian C. Bruce[2], Jan Wouters[1], and Tom Francart[1]

[1]ExpORL, Dept. of Neurosciences, KU Leuven. Herestraat 49, Bus 721, 3000 Leuven, Belgium.

[2]Dept. of Electrical and Computer Engineering, McMaster University. 1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada.

December 9, 2016

Abstract

Several filterbank-based metrics have been proposed to predict speech intelligibility (SI). However, they incorporate little knowledge of the auditory periphery. Neurogram-based metrics provide an alternative, incorporating knowledge of the physiology of hearing by using a mathematical model of the auditory nerve response. In this work, SI was assessed utilizing different filterbank-based metrics (the Speech Intelligibility Index and the Speech-based Envelope Power Spectrum Model) and neurogram-based metrics, using the biologically-inspired model of the auditory nerve proposed by Zilany et al., [2009, *The Journal of the Acoustical Society of America*, **126**(5), 2390–2412] as a front-end and the Neurogram Similarity Metric and Spectro Temporal Modulation Index as a back-end. Then, the correlations with behavioural scores were computed. Results showed that neurogram-based metrics representing the speech envelope showed higher correlations with the behavioural scores at a word level. At a per-phoneme level, it was found that phoneme transitions contribute to higher correlations between objective measures that use speech envelope information at the auditory periphery level and behavioural data. The presented framework could function as a useful tool for the validation and tuning of speech materials, as well as a benchmark for the development of speech processing algorithms.

I Introduction

Speech intelligibility (SI) is assessed using behavioural or objective measures. Usually, in the former, a group of participants are asked to listen to a stimulus under particular conditions and asked to register or identify what they heard (Miller, 2013). Such tests are used to characterise a patient's hearing and to evaluate the performance of new hearing devices. Furthermore, behavioural measures have allowed gathering valuable information on auditory perception. Objective measures are a complementary approach. They are also used to evaluate instruments' performance, since they present several advantages on their own. Their parameters can be set and tuned flexibly to investigate different conditions. Additionally, they can be obtained faster and in an automated way using a computer. Several objective measures have been used to predict SI. For the purposes of this paper, we will divide them into two groups: *filterbank-based* metrics and *neurogram-based* metrics.

On the one hand, filterbank-based metrics model the frequency selectivity of the auditory periphery by separating the speech signal into various frequency bands. There are several well established objective measures that have this working principle at their core, such as the Articulation Index (AI, French and Steinberg, 1947), the Speech Transmission Index (STI, Steeneken and Houtgast, 1980), the Speech Intelligibility Index (SII, ANSI, 1997, Sec. 2), and the Speech-based Envelope Power Spectrum Model, (sEPSM, Jørgensen and Dau, 2011, Sec. 2).

These metrics have been moderately successful in predicting SI of normal-hearing (NH) listeners under various conditions (e.g., Bradley, 1986; Jørgensen et al., 2013; Kryter, 1962; Pavlovic, 1987). However, their approach for modelling the auditory periphery can be thought of as simplistic, since they base their SI prediction on acoustic features within each band rather than on knowledge of any aspect of physiological processing performed by it. Furthermore, taking into account the anatomical and the physiological mechanisms underlying the auditory system can provide a better understanding of the different factors that affect SI of NH or hearing-impaired (HI) listeners. For example, incorporating biological information is essential when modelling effects of hearing impairment. A physiologically-based approach allows to incorporate different impairment conditions at various stages, e.g., sensorineural hearing loss due to damage to the inner hair cells (IHCs) or outer hair cells (OHCs).

On the other hand, neurogram-based metrics use mathematical models to mimic the physiological response of the auditory periphery. They represent neural activity as a function of characteristic frequency (CF) and time (i.e., the neurogram itself). These metrics try to predict SI with little influence of higher order processes (e.g., cognitive, linguistic, phonetic; Sidwell and Summerfield, 1986) by comparing neurograms of clean and corrupted speech. Since they take into account anatomy and physiological processes, we believe they can provide a better understanding of underlying factors in the auditory system that affect SI.

Different physiological models have been developed in the past to investigate different phenomena, such as pitch and timbre (Lyon and Shamma, 1996), the responses of high-spontaneous-rate auditory nerve (AN) fibers (Zhang et al., 2001), and to predict neural activity to speech (Bruce et al., 2003), for instance. Recently, Zilany et al. (2009) proposed a model of the physiological response at the AN level. It features a middle ear filter, two modes of basilar membrane excitation (including the known cochlear nonlinearities), and power-law dynamics and exponential adaptation for the synapse model. Its responses have been validated with physiological data over a wider dynamic range than previously existing models. This model has been successfully used to study a variety of auditory phenomena, such as neural adaptation to sound level (Zilany and Carney, 2010), sensory responses to musical consonance-dissonance (Bidelman and Heinz, 2011), overshoot adaptation (Jennings et al., 2011), masking release (Bruce et al., 2013), frequency selectivity (Jennings and Strickland, 2012), and neural coding of chimaeric speech (Heinz and Swaminathan, 2009). However, its use for assessing SI and the benefits of its biologically-inspired nature have been evaluated only in a limited number of studies so far. For example, Hines and Harte (2012) used it together with their Neurogram Similarity Metric (NSIM) to simulate performance intensity functions in quiet and in noise, which compared favourably with SII predictions of phoneme recognition in NH listeners. Zilany and Bruce (2007) used the model's previous version (Zilany and Bruce, 2006) together with a modified version of the Spectro Temporal Modulation Index (STMI, Elhilali et al.,

2003) and found good agreement between the model predictions and SI scores using filtered sentences at different presentation levels and with different levels of cochlear impairment. These two studies have been performed under different settings and conditions (e.g., speech material, noise, even with different versions of the model), making it hard to make a direct comparison of their results.

The objective of this study is to assess SI utilizing different objective measures and to compare their performance. We evaluate them in the same manner and under the same conditions, allowing for an understanding of their possibilities and shortcomings. For the neurogram-based metrics, we use the AN model proposed by Zilany et al. (2009) with parameters defined by Zilany et al. (2014) as a front end to generate neurograms (envelope – ENV, temporal fine structure – TFS, and early stage – ES) at different time scales. As a back end, we use the NSIM and STMI metrics as described by Hines and Harte (2012) and Elhilali et al. (2003), respectively. We compare their performance to two well-established filterbank-based metrics: the SII (ANSI, 1997) and the sEPSM (Jørgensen and Dau, 2011). Finally, we investigate whether these objective measures could predict behavioural scores or not by looking into their correlations with behavioural data. We hypothesize that the neurogram-based metrics will correlate at least as well as the filterbank-based metrics with the behavioural scores, since they encode physiological information that we think is important for speech understanding. We also believe that the neurogram based metrics that have an ENV-based front end will be correlated higher with the behavioural scores than those that do not,

since the literature suggests that the ENV component of speech has a large contribution to its perception (Drullman, 1995; Shannon et al., 1995; Swaminathan and Heinz, 2012).

The manuscript is organized as follows. Section II provides the technical background of the objective measures used in this work. Section III describes how the study was conducted. Section IV presents the obtained results, which are further discussed in Sec. V. Section VI closes the paper with our overall conclusions.

II Background

A Neurogram-based Metrics

1 AN Model

The model proposed by Zilany et al. (2014, 2009) is capable of reproducing response properties of AN fibers. It is comprised of various modules, each mimicking a particular function of the auditory periphery. First, the stimulus is passed through a filter emulating the middle ear. The output is then passed through a signal path and a control path. The former simulates the behaviour of the OHC-controlled filtering properties of the basilar membrane (BM) in the cochlea and the transduction properties of the IHCs by a succession of non-linear and low-pass filters. The latter simulates the function of the OHCs in controlling BM filtering. The output of the control path feeds back into itself and into the signal path, as well. The IHCs output then goes through an IHC-AN synapse module with two power-law adaptation paths, accounting for slow and fast adaptations.

2 Neurograms

For each input, the AN model produces three different neurograms: the ENV, the TFS, and the ES neurograms, which are generally explained as follows. Further details on their implementation are given in Sec. 1.

The ENV and TFS neurograms[1] allow studying the neural response at different time resolutions (Hines and Harte, 2010, 2012). The ENV neurogram represents smoothed (averaged) discharge rate using a bin size of 6.4 ms. Thus, only slow temporal modulations related to the ENV are available. On the other hand, the TFS neurogram retains spiking information and phase-locking related events (Young, 2008). The TFS neurogram uses a bin size of 0.16 ms. Both of them are obtained by convolving them with a Hamming window of 128 and 32 samples, respectively, with 50 % overlap.

The ES neurogram explicitly encodes temporal envelope modulations due to the interplay of the spectral components in each band (Elhilali et al., 2003). In this case, the neural activity was binned into time bins of 8 ms. It was obtained by convolving it with a rectangular window of 2 samples with a 50 % overlap.

3 Similarity Metrics

In order to obtain a measure of SI for each speech token, two different neurograms are computed: a reference neurogram r (which receives the speech token in quiet as an input) and a degraded neurogram d (which receives the speech token in noise as an input). Then, the similarity between the two neurograms can be calculated using different metrics, in our case the NSIM and the STMI.

NSIM The NSIM is a simplified version of the Structural Similarity Index (SSIM, Wang et al., 2004). It considers the neurograms as images and quantifies their similarity as a function of their luminance l (comparing the mean values across both images) and their structure s (equivalent to their correlation coefficient), as given by Eq. 1 and 2:

$$\text{NSIM}(r, d) = l(r, d) \cdot s(r, d) \tag{1}$$

$$= \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} \cdot \frac{\sigma_{rd} + C_2}{\sigma_r\sigma_d + C_2} \tag{2}$$

In the latter equation, μ and σ represent the mean intensity and standard deviation, respectively, of their corresponding neurograms, while σ_{rd} represents the covariance between both of them. Each factor contains constant values $C_1 = 0.01L$ and $C_2 = (0.03L)^2$ (where L is the intensity range). Although these have little effect on the metric value itself, they are useful to prevent instabilities at boundary conditions.

STMI The STMI is a measure of speech modulation integrity. It quantifies the degradation of the speech modulations in the temporal and the spectral dimensions jointly due to the addition of noise (regardless of its nature) or processing of the speech signal itself. It takes the AN activity and projects it into a higher, more central level at the primary auditory cortex. To do so, it applies a bank of modulation selective filters (which resemble those described in the mammalian central auditory system, Chi et al., 1999; Wang and Shamma, 1995) to the input

neurograms. The former consists of 9 temporal filters ranging from slow to fast rates and of 11 spectral filters ranging from narrow to broad scales. This yields a 4D representation of the activity at a central (i.e., cortical) level: time, frequency, temporal rate, and spectral scale. Then, in order to only extract temporal and spectral modulations, the cortical representation is adjusted by subtracting the ES neurogram of a base signal (i.e., a signal with the same long-term spectrum but randomized phase). Finally, the STMI is calculated between the reference cortical representation R of the neurogram r (corresponding to the word/phoneme in quiet) and the cortical representation D of the neurogram d (corresponding to the word-phoneme in noise) using Eq. 3:

$$\mathrm{STMI}(R, D) = 1 - \frac{||R - D||^2}{||R||^2} \tag{3}$$

B Filterbank-based Metrics

The SII and the sEPSM are two well established filterbank-based SI metrics. The SII is a well known measure of SI that uses a relatively simple filterbank to model the ear's frequency selectivity. The sEPSM also includes such a filterbank, but it goes one step further and uses an additional (modulation) filterbank.

1 SII

The SII computes the average amount of useful speech information that is available to the listener (ANSI, 1997). Mathematically, it is given by Eq. 4.

$$\mathrm{SII} = \sum_{i=1}^{n} \mathrm{FIF}_i A_i \tag{4}$$

It receives as an input the *clean* speech and the noise signal (Fig. 1). First, it partitions the inputs into n individual frequency bands. These can be one-third-octave bands, octave bands, or critical bands. Next, for each band, its audibility A (i.e., the proportion of audible speech cues that are audible to the listener) is calculated. A is simply based on the level of speech relative to the level of noise. For its computation, the spectrum level of noise is subtracted from the spectrum level of speech in each band. Then, correction factors (designed to account for distortion due to high presentation levels and upward spread of masking) are applied. Lastly, the SNR is computed and normalized between 0 and 1 (assuming a dynamic range of speech of 30 dB). Next, A is multiplied by the

band frequency importance function (FIF), which determines the contribution of different frequency regions to speech recognition. The FIF depends on the type of speech material and presentation level. The sum of these values across all bands is approximately equal to 1. Finally, these values are summed across the different frequency bands, yielding a single SII value (Hornsby, 2004).

2 sEPSM

The sEPSM (Jørgensen and Dau, 2011) is an extension of the EPSM, originally proposed by Dau et al. (1999) and Ewert and Dau (2000). It receives as an input the *degraded* speech and the noise signal (Fig. 1). First, it passes each input through a gammatone filterbank. Then, the envelope of each channel is extracted using the Hilbert transform. The resulting envelope is input to a modulation filterbank and the power of the filtered envelope computed, resulting in $P_{\text{env}\,S+N}$ and $P_{\text{env}\,N}$ for the degraded speech signal and the noise signal, per channel, respectively. After that, the envelope SNR of a channel i ($\text{SNR}_{\text{env}\,i}$) is computed using Eq. 5. Next, the SNR_{env} of all n channels is combined into a single overall value, using Eq. 6.

$$\text{SNR}_{\text{env}\,i} = \frac{P_{\text{env}\,S+N} - P_{\text{env}\,N}}{P_{\text{env}\,N}} \tag{5}$$

$$\text{SNR}_{\text{env}} = \sqrt{\sum_{i=1}^{n} (\text{SNR}_{\text{env}\,i})^2} \tag{6}$$

Following, the overall $\mathrm{SNR}_{\mathrm{env}}$ is transformed to a sensitivity index d' of an ideal observer using Eq. 7, where k and q are speech-material-dependent parameters.

$$d' = k \cdot \mathrm{SNR}_{\mathrm{env}}^{q} \tag{7}$$

Finally, d' is converted into the probability of the ideal observer of correctly recognizing the speech item P_{correct} using the mAFC model proposed by Green and Birdsall (1964). This model compares the input speech element with a set of m previously stored alternatives. Then, it chooses the most similar one, x_S. x_S is a random variable with mean d' and variance σ_S^2 (which is related to the redundancy of the speech material). The remaining $m-1$ items are considered to be noise. Of these, the one that has the largest similarity with the input speech element is chosen as x_N. The latter is also a random variable with mean μ_N and variance σ_N^2. P_{correct} is calculated from the difference distribution of x_S and x_N, as given by Eq. 8. Φ stands for the cumulative normal distribution.

$$P_{\mathrm{correct}} = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right) \tag{8}$$

III Materials & Methods

An overview of the materials and methods used is provided in Fig. 1.

A Speech Material

The *Leuven Intelligibility Peutertest* (Lilliput) speech material was used in this experiment. The full corpus consists of 378 meaningful Flemish CVC words uttered by a female speaker. However, in order to improve its homogeneity, we selected words that were within one standard deviation around the mean of their average speech recognition threshold (SRT) for adults (i.e., words with an SRT within the -9.8 ± 2.9 dB range). Additionally, given the time needed for further segmentation (Sec. 1), a subset of 65 randomly-picked words was finally chosen.

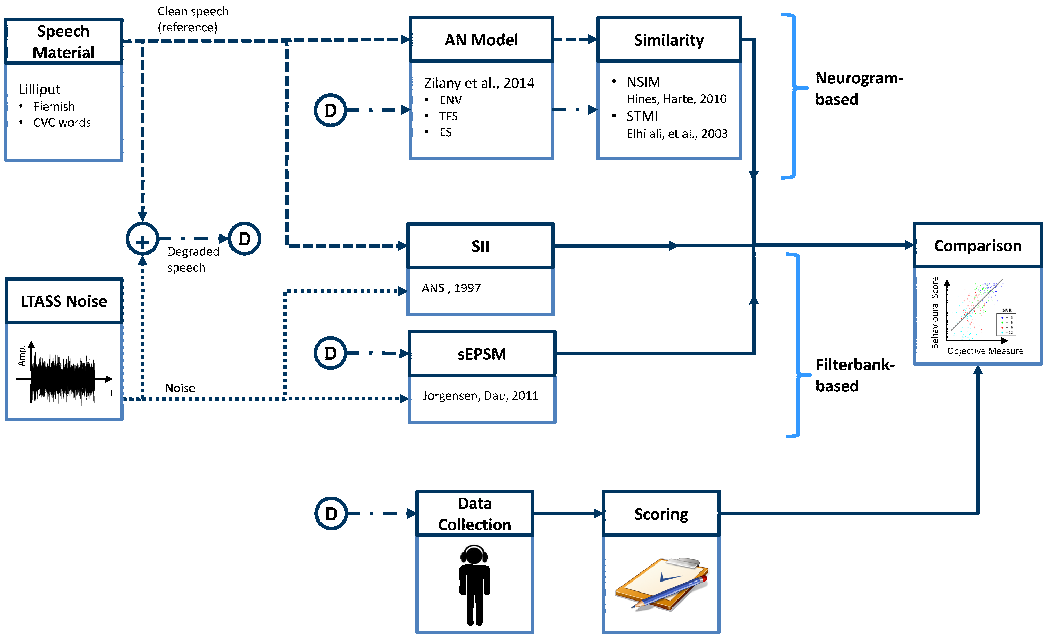

Figure 1

Stimuli were combined with the accompanying long-term averaged speech-shaped (LTASS) noise at five different SNRs: from 0 to -12 dB in steps of 3 dB. These degraded audio files were the target material for the objective measures (Sec. B) and the behavioural measurements (Sec. C).

1 Segmentation

On one hand, we were interested in studying perception at a *word level* (i.e., with no sentence context), since it has been shown that sentence context has an influence on word perception (Boothroyd and Nittrouer, 1988), which cannot be easily modelled. On the other hand, we were interested in studying perception at a *phoneme level*, given that phoneme scores present a reduced variability and thus exhibit greater test-retest reliability (Gelfand, 1998).

Thus, for the latter each audio file was manually segmented into phonemes. Two different kind of segmentations were done. In the first, the segmented audio only included the sound of its corresponding phoneme, yielding segments C_1, V, and C_2 (*pure-phoneme segments*). In the second, the segmented audio additionally included the transition to and/or from the preceding/succeeding phoneme, yielding segments Cv, cVc, and vC (*transitions-included segments*). The phoneme limits were determined using Praat 5.3.16 (Boersma and Weenink, 2014). These were delimited by visual inspection of the time signal and the spectrogram, together with auditory inspection. Figure 2 shows an example segmentation of the word *bot*.

B Objective Measures

1 Neurogram-based Metrics

The degraded and clean (reference) signal were fed to the mathematical model of the AN proposed by Zilany et al. (2014, 2009) with parameters set to simulate NH listeners. Each input produced an ENV, a TFS, and an ES neurogram (Sec. 2).

All neurograms depicted the average response of 50 AN fibers at each CF with different spontaneous rates: high (100 spikes/s), medium (5 spikes/s), and low (0.1 spikes/s), with weights of 0.6, 0.2, and 0.2, respectively, corresponding to the distribution observed in animals (Zilany and Bruce, 2007). The ENV and TFS

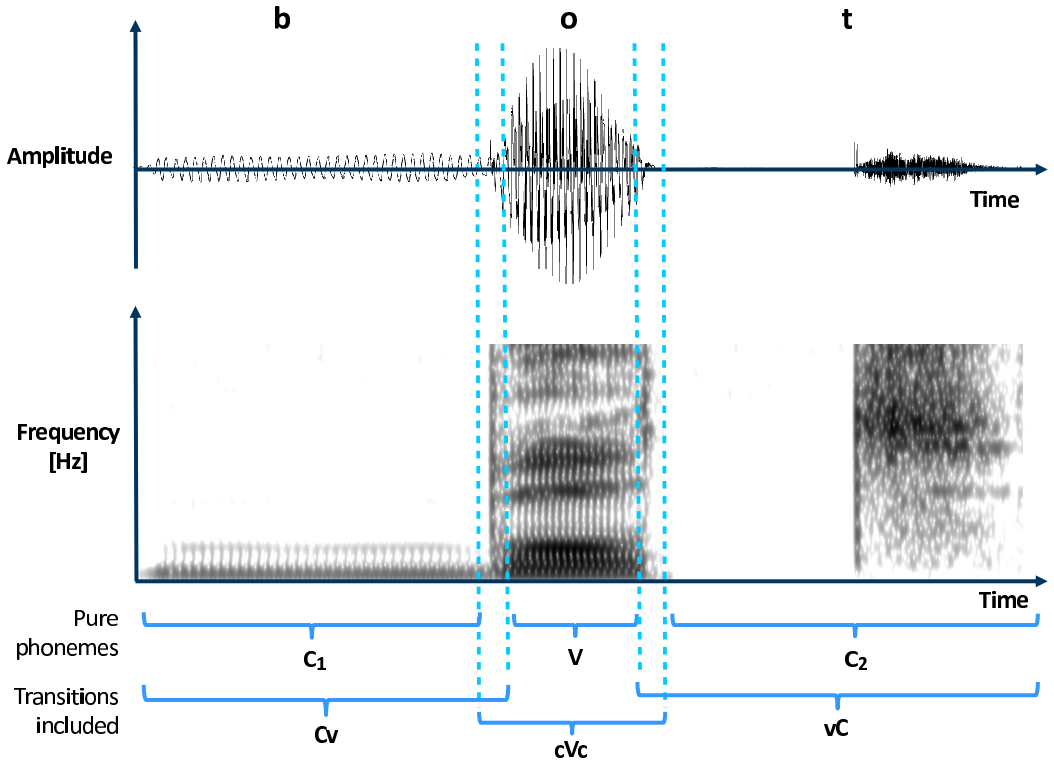

Figure 2

neurograms considered 30 CFs logarithmically spaced from 250 to 8000 Hz (Hines and Harte, 2010, 2012). The ES neurogram considered 128 CFs logarithmically spaced from 180 to 7000 Hz (Elhilali et al., 2003). Figure 3 shows example ENV and TFS neurograms in quiet and at different SNRs for the word *bot*. Figure 4 shows example ES neurograms and cortical representations for the same word under the same conditions.

After that, the deterioration from the reference speech token r to the token of the degraded stimulus d was quantified using the NSIM and the STMI. Specifically, the NSIM metric was applied to the ENV and TFS neurograms (Hines and Harte, 2010, 2012). The STMI metric was applied to the ES neurogram (Elhilali et al., 2003). Additionally, we applied the STMI to the ENV and TFS neurograms, as well, in order to explore the results of projecting the information of such neurograms to a higher (more central) level. Care was taken to make sure that the STMI modulation filters covered the correct range of spectral modulation scales for the corresponding neurogram's CFs. Thus, the STMI spectral filters ranged from 0.25 to 8 cycles/oct, while the STMI ENV and STMI TFS spectral filters ranged from 0.25 to 2 cycles/oct. The temporal filters in both cases went from 2 to 32 Hz. For all metrics, the value for each word/phoneme was averaged across participants for each SNR condition. Lastly, a straight line was fitted through these points.

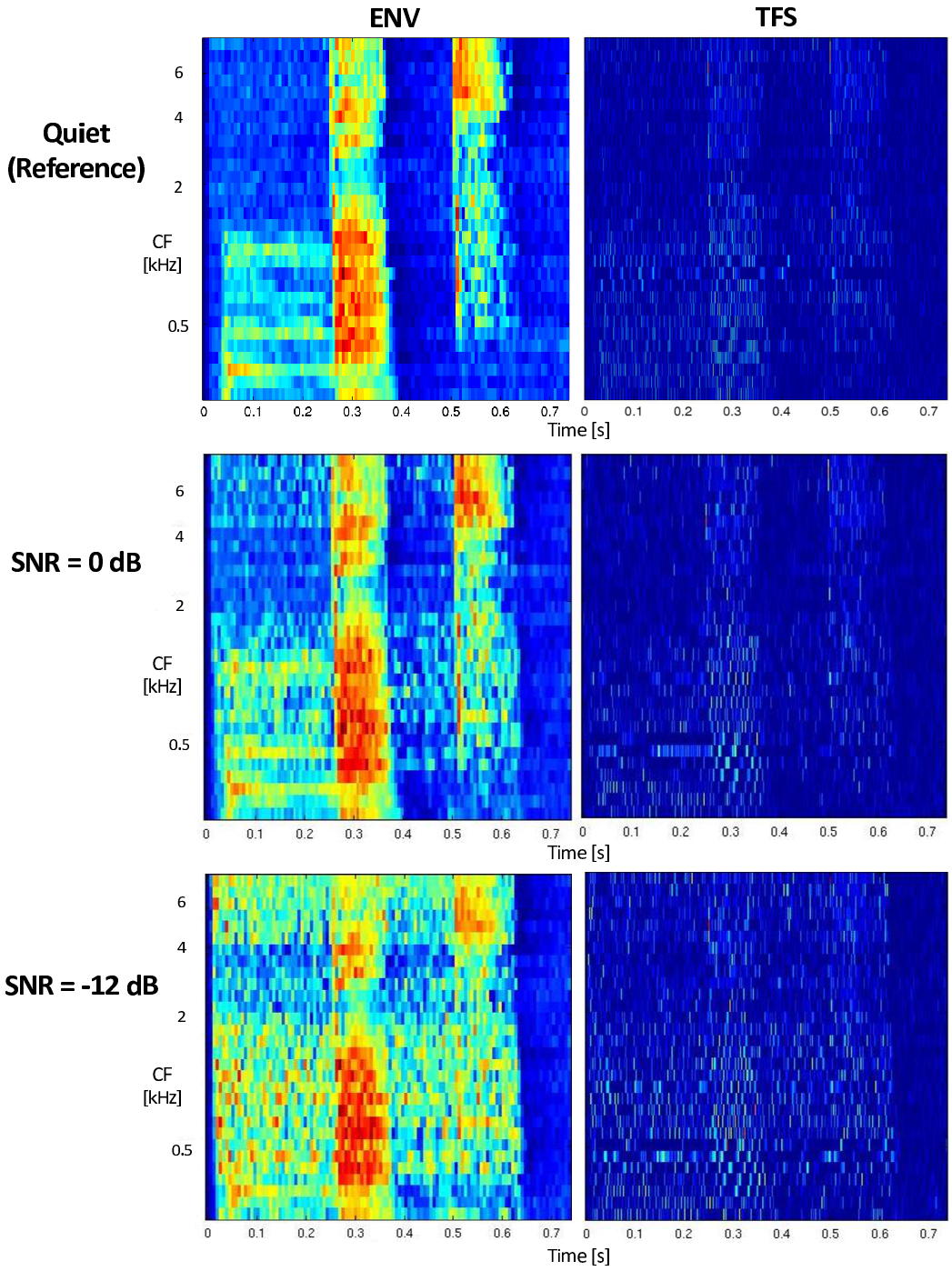
ENV
TFS
Quiet
(Reference)
SNR = 0 dB
SNR = -12 dB
CF
[kHz]
6
4
2
0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time [s]

Figure 3

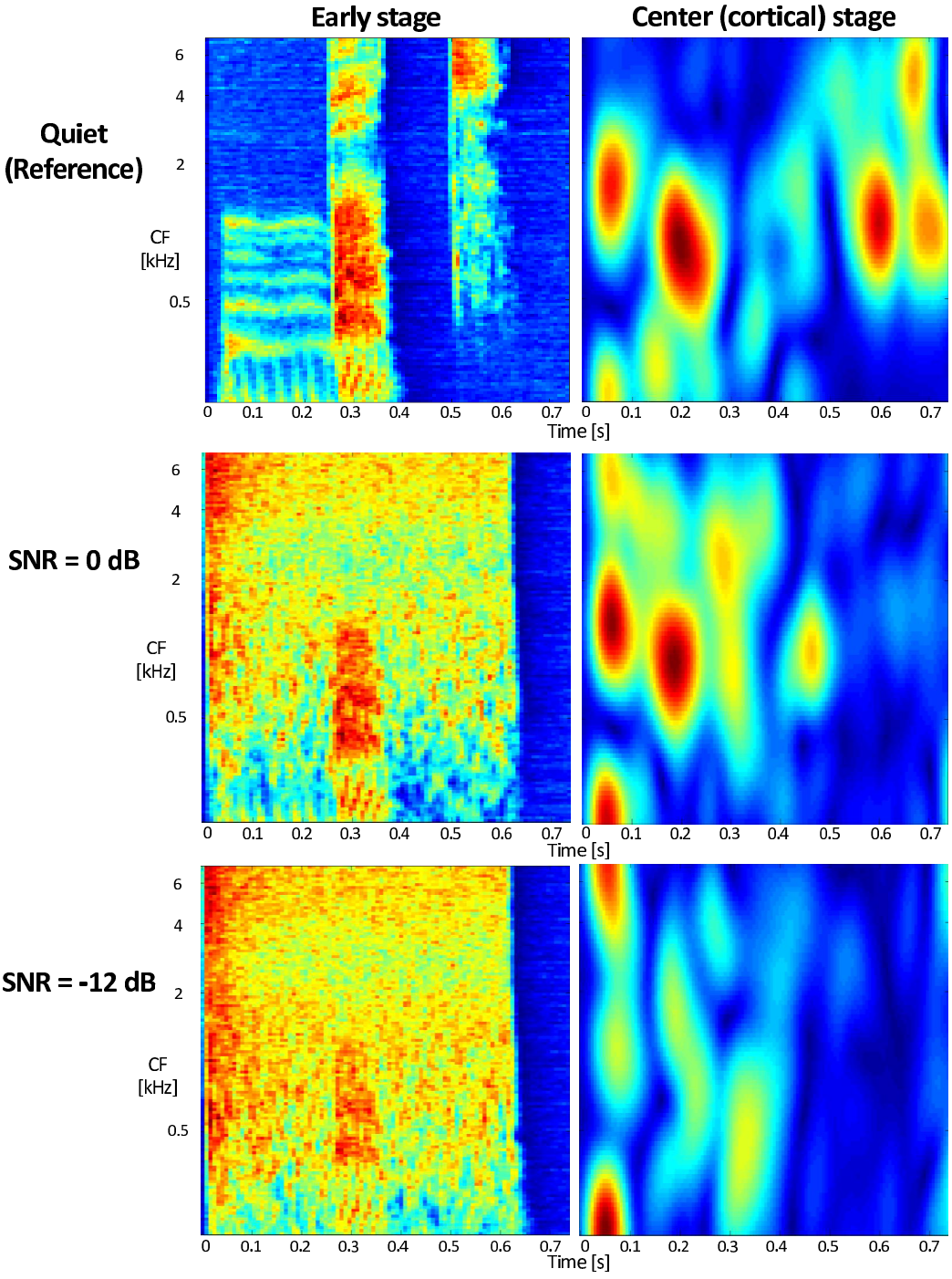
Early stage
Center (cortical) stage
Quiet
(Reference)
SNR = 0 dB
SNR = -12 dB
CF
[kHz]
6
4
2
0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time [s]

Figure 4

2 Filterbank-based Metrics

In order to be make our results comparable with those of previous studies (e.g., Hines and Harte, 2012; Hossain et al., 2016; Mamun et al., 2015), we chose to partition the input spectra into one-third octave bands for the computation of the SII. Furthermore, we used the SPIN FIF (ANSI, 1997, Table B.2).

In the case of the sEPSM, we used values of 0.275 and 0.315 for the ideal observer parameters k and q, respectively. These were empirically obtained by minimizing the root mean square error (RMSE) between the model predictions and the psychometric function of the behavioural data. In order to reduce over-fitting, we obtained these values using only one-third of the available data. To convert d' into P_{correct}, m was assumed to be 8,000 (the size of a person's active vocabulary, Müsch and Buus, 2001), since we used an open set.

C Behavioural Measurements

1 Data Collection

Twenty participants (5 males, 15 females, mean age 20.75 ± 1.48 years old) volunteered for this study. They were tested and confirmed to have NH. All of them were native Flemish speakers and provided written informed consent. The study was approved by the local ethics committee.

Measurements were performed in a sound booth. Words were presented to the participant at a fixed speech level of 65 dB sound pressure level (SPL) using APEX 3 (Francart et al., 2008). These were routed from a computer via an external RME Fireface sound card to Sennheiser HD 250 Linear II headphones. Words were presented randomly across SNRs. For each trial, participants were instructed to listen to the stimulus and to type the word they thought they had heard into the computer. Each word was presented once for each SNR condition.

2 Scoring

For each trial, two different behavioural scores were obtained: a *phoneme score* (at a word level) and a *per-phoneme score* (at an individual phoneme level).

The phoneme score was assigned by the experimenter to the participant's answer depending on the number of phonemes that were correct compared to the original word. For example, if the presented word was *bot* and the participant's answer was *bol*, a phoneme score of 2 was given. Phoneme scores for each word

and SNR were averaged across participants.

Per-phoneme (i.e., individual phoneme) scores can be obtained in two ways. On one hand, they can be computed *a posteriori* by comparing the participant's response to the original word and assigning a 1 or a 0, indicating if each phoneme was correct or not, respectively. On the other hand, the phonemes can be presented individually to the participants in a controlled context (e.g., aCa or pVp). For this study, we preferred the former approach, since in this case the phonemes are presented in a more natural context and in a more similar to way to real-life realizations. Although there might be some lexical influence (Ganong, 1980), given the token size we expect that most (if not all) contextual effects are present, thus limiting certain lexical biases in the responses. Furthermore, the speech material used here consists of very short words (CVC), which minimizes such effects even more. Following the previous example, if the presented word was *bot* and the participant's answer was *bol*, a per-phoneme score of [1 1 0] was calculated. Automatic computation of the per-phoneme scores was done using the algorithm proposed by Francart et al. (2009). The algorithm's performance was evaluated by comparing it with the annotated phoneme score (in the end, the phoneme score is the sum of the three per-phoneme scores), with success in 94 % of the cases. The rest of the instances were evaluated manually. Per-phoneme scores were also averaged across participants.

D Comparison

We observed a ceiling effect in the behavioural scores, particularly at the condition of SNR = 0 dB. Since this cannot be modelled by most of our objective measures, these data points were not considered for further analysis.

Scatter plots of the scores versus the metrics across different SNRs were made at word and phoneme level. We computed a simple linear regression model in each case. Pearson correlation coefficients were computed between the behavioural and the objective variables. For their comparison, we used William's test (Williams and Williams, 1959) with Bonferroni correction. Additionally, the goodness of the linear regression was evaluated using the F-ratio, which quantifies the improvement of the model prediction compared to the level of the model inaccuracy (Field et al., 2012).

Finally, we evaluated what unique proportion of the variance was explained by the different objective measures. Using a hierarchical predictor selection approach, we chose the metrics of each of the objective measures groups (neurogram-based and filterbank-based) with the highest correlations and used these to generate multivariate linear regression models. Then, we computed the R^2 values for each case and looked at the difference between these values and those obtained in the simple linear regression.

IV Results

The results for the objective measures at word level are shown in Fig. 5 and 6. The former shows the metrics for 15 randomly picked (sample) words. The latter shows the boxplot of these metrics, as well as its average across the whole set of 65 words. The dotted line corresponds to the fitted straight line. Overall, we found a directly proportional relation between the objective metrics and SNR at a word and phoneme level.

Figure 7 shows the scatter plots of phoneme scores vs different metrics averaged across participants at a word level. In this case, each point corresponds to a word. For each metric, the Pearson correlation was calculated between the phoneme score and the objective measure and is shown on the bottom right corner in each plot, together with its Bonferroni corrected p-value. ENV-based metrics had a stronger correlation with behavioural scores than TFS-based metrics: the NSIM ENV correlation (0.73) was significantly higher ($p < 0.01$) than the NSIM TFS correlation (0.67); the STMI ENV presented a 0.69 correlation, which is significantly higher ($p < 0.001$) to its TFS counterpart (0.24). In the case of the filterbank-based measures, the SII and sEPSM metrics had a correlation of 0.61 and 0.35, respectively, with $p < 0.001$ for both of them. Overall, the NSIM ENV and STMI ENV showed the strongest correlations of all the metrics ($p < 0.05$) with no significant differences between them ($p > 0.05$). Additionally, we computed the F-ratios for the linear regression of each of the different ob-

jective measure (also shown in their corresponding scatter plot together with its p-value). Similar to the Pearson correlations, the NSIM ENV, NSIM TFS, and STMI ENV metrics had the largest F-ratios (275, 225.9, and 219.9, respectively, all with $p < 0.001$). In the case of the filterbank-based measures, the SII and sEPSM metrics had F-ratios of 120.6 and 28.4, respectively, with $p < 0.001$ for both of them.

Based on these results, we chose the NSIM ENV and STMI ENV metrics from the neurogram-based group and the SII from the filterbank-based group as predictors for the multivariate linear regression. Analysing the R^2 values allowed us to quantify the amount of unique proportion of variance explained by the different objective measures. For instance, the regression model that started with the SII had an R^2 value of only 0.37. When we incorporated either the NSIM ENV or STMI ENV metrics into it, the R^2 value significantly increased by 0.18 (ANOVA, $F(1, 259) = 81.51$, $p < 0.001$) and 0.16 (ANOVA, $F(1, 259) = 70.95$, $p < 0.001$), respectively. Going the other way around, when the regression model started with the NSIM ENV or STMI ENV metrics, the R^2 values were of 0.54 and 0.47, respectively (which are already higher than in the SII model case). When we incorporated the SII metric, the R^2 values significantly increased by barely 0.01 (ANOVA, $F(1, 259) = 4.48$, $p < 0.05$) and 0.06 (ANOVA, $F(1, 259) = 25.38$, $p < 0.001$), respectively. A summary of these results is shown in in Table 1.

Table 1: Results of the (multivariate) linear regression models

Predictors (objective metrics)			R^2	F-statistic	DoF	p-value
NSIM ENV	STMI ENV	SII				
✓			0.54	199.40	260	< 0.001
	✓		0.47	146.10	260	< 0.001
		✓	0.37	86.22	260	< 0.001
✓		✓	0.55	103.90	260	< 0.001
	✓	✓	0.53	96.02	260	< 0.001

Additionally, Fig. 8 shows the Pearson correlation between the per-phoneme scores vs different metrics for pure phoneme segments and for phoneme segments that include transitions. In this case, weak to moderate correlations were mostly found. We were interested in studying the influence of including transitions on the correlation between the objective measures and the behavioural per-phoneme scores. We found significantly lower correlations of C_1, V, and C_2 compared to their counterparts Cv, cVc, and vC in the NSIM ENV ($p < 0.05$, $p < 0.001$, and $p < 0.05$, respectively); significantly lower correlation of C_2 compared to vC in the STMI TFS ($p < 0.05$); significantly lower correlation of C_1 compared to Cv in the SII ($p < 0.05$). The rest of the metrics did not show significant differences.

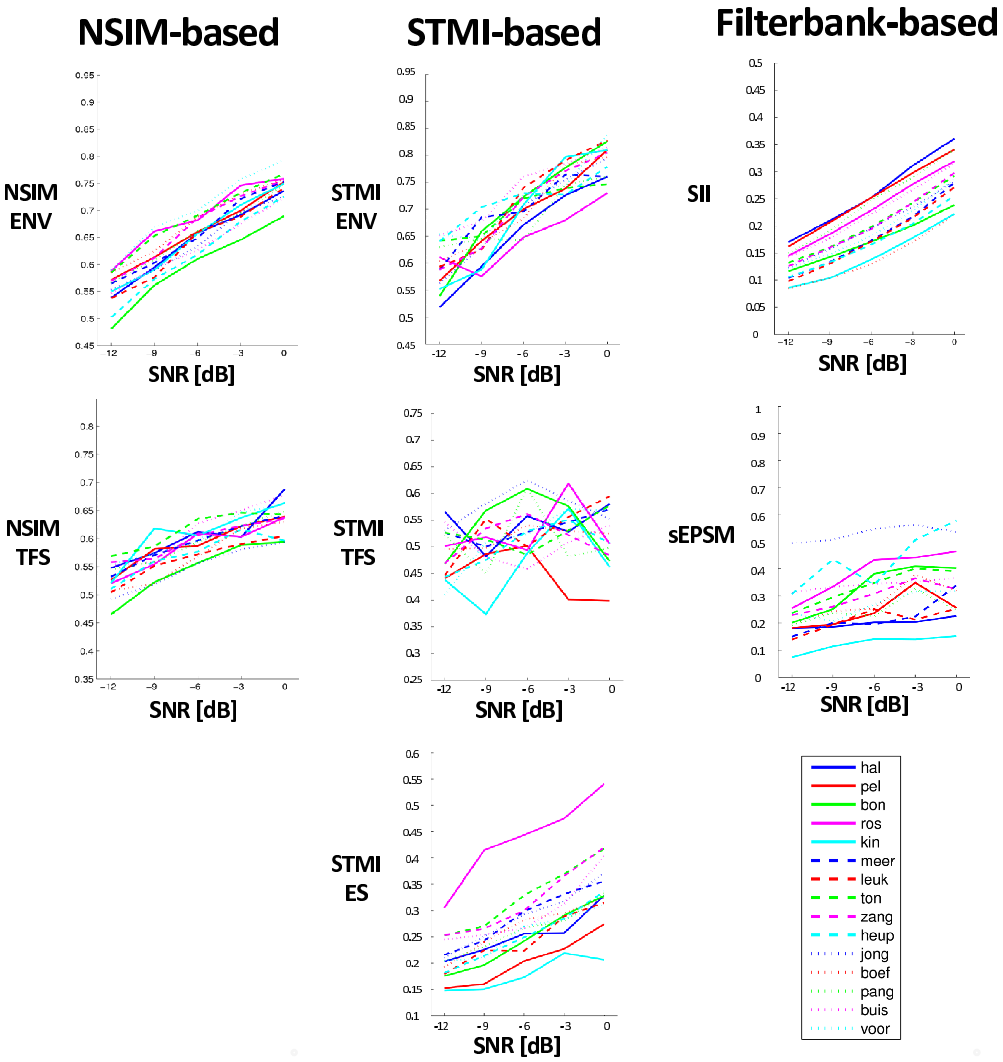
NSIM-based
STMI-based
Filterbank-based
NSIM ENV
STMI ENV
SII
NSIM TFS
STMI TFS
sEPSM
STMI ES
SNR [dB]
hal
pel
bon
ros
kin
meer
leuk
ton
zang
heup
jong
boef
pang
buis
voor

Figure 5

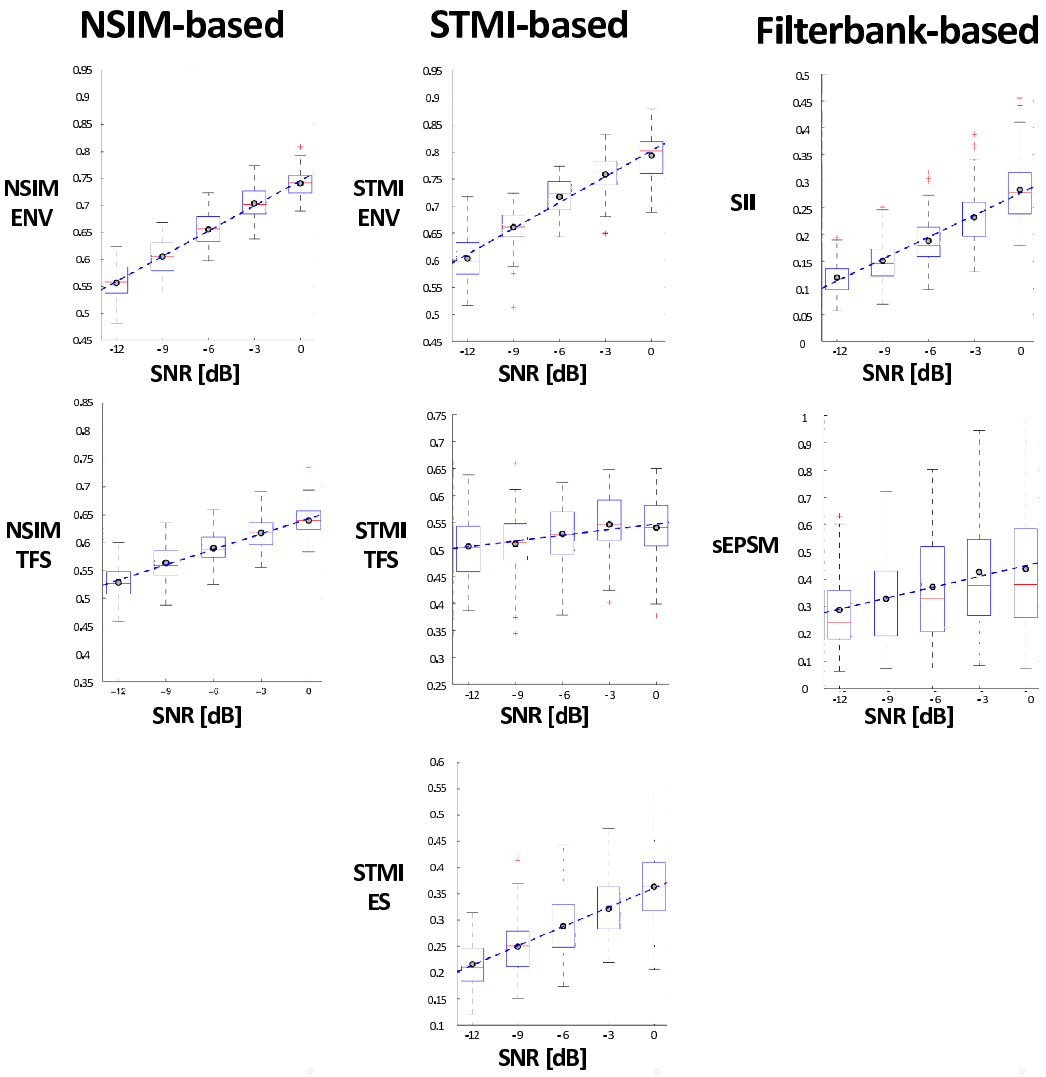
NSIM-based
STMI-based
Filterbank-based
NSIM ENV
STMI ENV
SII
NSIM TFS
STMI TFS
sEPSM
STMI ES
SNR [dB]

Figure 6

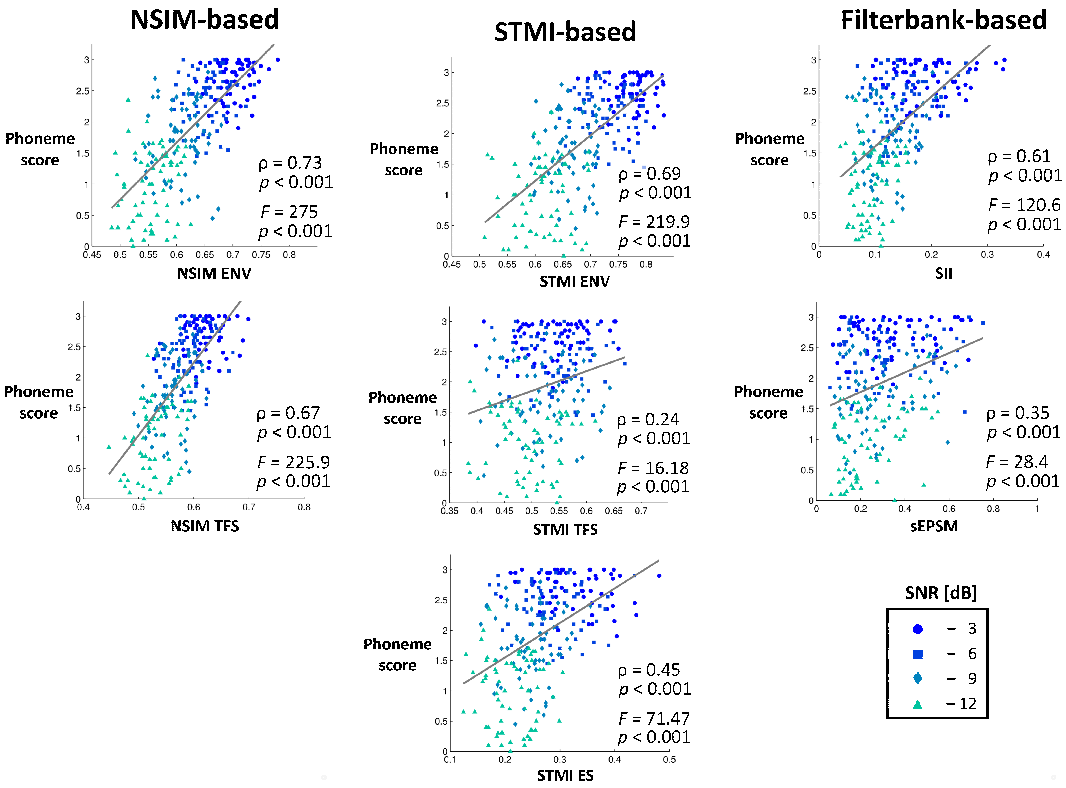
NSIM-based
STMI-based
Filterbank-based
Phoneme score
ρ = 0.73
p < 0.001
F = 275
p < 0.001
NSIM ENV
ρ = 0.69
p < 0.001
F = 219.9
p < 0.001
STMI ENV
ρ = 0.61
p < 0.001
F = 120.6
p < 0.001
SII
ρ = 0.67
p < 0.001
F = 225.9
p < 0.001
NSIM TFS
ρ = 0.24
p < 0.001
F = 16.18
p < 0.001
STMI TFS
ρ = 0.35
p < 0.001
F = 28.4
p < 0.001
sEPSM
ρ = 0.45
p < 0.001
F = 71.47
p < 0.001
STMI ES
SNR [dB]
− 3
− 6
− 9
− 12

Figure 7

V Discussion

In this study, we assessed SI using different neurogram-based and filterbank-based objective measures. Then, we calculated the correlations between these and behavioural scores of NH listeners.

Figures 5 and 6 show how metric values increase together with SNR. This is due to the fact that noise is random and spurious information which shows in the neurograms as activity (Fig. 3), but that actually diminishes the metric.

At a word level, the strongest correlations were found between the behavioural scores and the ENV-based metrics NSIM ENV and STMI ENV, showing correlations of 0.73 and 0.69, respectively (with no significant difference between them, $p > 0.05$). These are significantly higher than their TFS counterparts: NSIM TFS with a value of 0.67 and STMI TFS with a value of 0.24 ($p < 0.05$ and $p < 0.001$, respectively); they were also significantly higher than those found

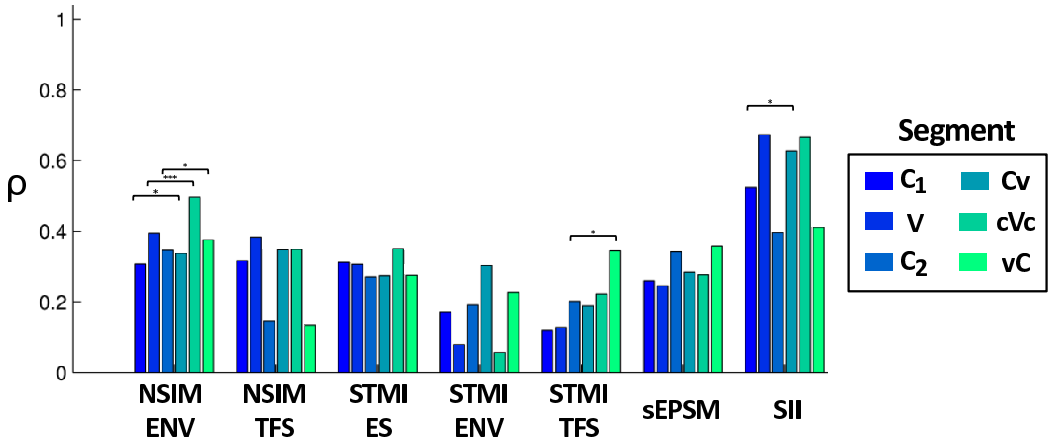

Figure 8

using filterbank-based metrics: SII with a value of 0.61 and sEPSM with a value of 0.35 ($p < 0.01$ and $p < 0.001$, respectively). This suggests that the ENV component of speech (as represented by the ENV neurograms) has a stronger correlation with the behavioural scores than the TFS component. This goes in line with what is reported in literature, which suggests that the ENV component of speech has a larger contribution than the TFS for its perception (e.g., Drullman, 1995; Shannon et al., 1995; Smith et al., 2002; Swaminathan and Heinz, 2012). Furthermore, it hints that neurogram-based metrics correlate as much (or higher) than filterbank-based metrics. This could be due to the fact that the former incorporate physiological aspects of the auditory system as part of the AN model. The STMI TFS metric had the lowest correlation of all (0.24), while also showing large variance across different listeners. This suggests that the forenamed objective measure is not a reliable predictor for SI. We hypothesize this is because TFS is lost in the base signal of the STMI. Furthermore, the information of the TFS is scattered and sparse and thus incapable of fully reflecting the modulations of the original speech token. These modulations are the base of the cortical representation of the STMI metric, hence it fails to provide a representation of the original input. The computed F-ratios show that of the proposed objective measures, the linear regression models of the NSIM ENV ($F = 275$, $p < 0.001$), NSIM TFS ($F = 225.9$, $p < 0.001$), and STMI ENV ($F = 219.9$, $p < 0.001$) metrics fit the data the best, since their large values reflect smaller differences between the model's predictions and the observed data. On the contrary, the

STMI TFS model ($F = 16.18$, $p < 0.001$) had the smallest F-ratio, reinforcing the idea that in the presented framework, this metric is a poor predictor of SI.

The simple linear regressions showed that from the NSIM ENV, STMI ENV (both neurogram-based), and the SII (filterbank-based) metrics, the latter had the smallest R^2 value. Furthermore, the R^2 values of the multivariate linear regressions showed the smallest improvement when the SII was incorporated. This suggests that (ENV) neurogram-based metrics are able to account for a larger proportion of the variance than filterbank-based metrics (SII, in this case), endorsing their value for SI prediction in the presented framework.

At a phoneme level, we found significantly lower correlations of C_1, V, and C_2 compared to their counterparts Cv ($p < 0.05$), cVc ($p < 0.001$), and vC ($p < 0.05$) in the NSIM ENV case. The fact that this was the only metric that was consistently sensitive to phoneme transitions could suggest that that these have a larger impact on intelligibility on the ENV component of speech (where they can be captured), rather than in the TFS. Furthermore, this could also hint that phoneme transitions have a larger impact on the information at the AN level, rather than at a higher level (e.g., cortical representation). However, these results have to be handled with care, since there is no clear agreement in the literature regarding the impact of transitions in different speech intelligibility tasks. On one hand, it has been suggested that transitions have an important impact on phoneme identification. Jenkins et al. (1994) found that the onset and offset of vowels in /dVd/ syllables was enough to identify the syllable. Strange and Bohn (1998) showed

that perceptual differentiation of German vowels is dependent on spectral information contained in transition cues (e.g., onsets and offsets). On the other hand, the opposite hypothesis has also been proposed. Cole et al. (1996) found that the location of the segment boundaries (i.e., the inclusion or exclusion of phoneme transitions) does not have a strong impact in consonants or vowels intelligibility. Using an entropy-based approach, Stilp and Kluender (2010) found that intelligibility patterns of replaced vowel-consonant and consonant-vowel transitions were indistinguishable from that of vowels. Fogerty and Kewley-Port (2007, 2009) concluded that for speech tokens with vowels present, the information in the transitions does not contribute to intelligibility since it might be redundant with that found at the center of the vowel (they agreed, though, that for consonant-only speech tokens, information in the transitions did provide a perceptual benefit). Lee and Kewley-Port (2009) found that different transitional information had a similar effect in speech intelligibility for young NH and elderly HI listeners.

Although comparing the metrics' performance between the word and phoneme level cases was not the main objective of this analysis, it is still worth mentioning a few things. Overall, we found stronger correlations in the former compared to the latter case. We hypothesize this could be due to a variety of reasons, depending on the metric. For the NSIM-based metrics, we think that using shorter neurograms reduces the amount of useful information available for the metric when comparing the reference neurogram with the degraded one. In the case of the STMI-based metrics, additionally we believe that the length of the

phoneme segments is not long enough to be correctly captured by all the filters of the temporal filter bank.

Finally, even though physiological-inspired frameworks (such as the one presented here) are successful in predicting SI, they still have a few shortcomings that are worth pointing out. Speech perception is a very complex process. The mapping of the speech signal along the auditory pathway is an intricate mechanism that is not yet fully understood. Studying SI using different approaches (Allen, 2005) is the first step towards a better comprehension of the various processes involved (e.g., the study of the learning component of speech understanding has served as the base of the development of automatic speech recognition systems, Benzeghiba et al., 2007). Additionally, current biologically-inspired models have been validated mostly using animal data. Their translation to human auditory processes rely on several assumptions, many of which still need to be confirmed by further physiological studies.

VI Conclusions

In this work, we investigated the correlation of different objective metrics with behavioural scores, with special emphasis on neurogram-based metrics that use the AN model proposed by Zilany et al. (2014, 2009) as a front end.

The relation between the objective measures (when averaged across participants and across words) and SNR can be explained by fitting a straight line. Furthermore, we found significantly stronger correlations between behavioural measurements and the ENV based objective measures at a word level. This goes in line with the usefulness of the ENV for behavioural perception, with the NSIM ENV and STMI ENV presenting the strongest ones and being able to explain the largest variance proportion. Besides, these objective measures present a few more advantages over the rest. Since they are based on the responses of the AN, they inherently incorporate physiological information. This provides a more transparent approach to understanding the processes occurring in the auditory system. Furthermore, thanks to the versatility of the AN model, it allows to incorporate biologically the effects of hearing loss due to damage to the IHCs, OHCs or both, something that is not possible to do straightforwardly using one of the filterbank-based approaches. Additionally, the latter rely heavily on calibration of their parameters for different cases (e.g., speech material, noise conditions), which is not only hard to achieve, but it can also lead to overfitting. At a per-phoneme level, we found that the NSIM ENV was consistently sensitive to the

phoneme transitions for C_1, V, and C_2. Lastly, we could not find evidence that simulating processes at a central level using the current approach (i.e., applying a cortical model on top of a peripheral representation) provides extra benefit over the information already available at the AN.

A framework like the one presented here could have different applications. It could function as a tool for the validation and tuning of speech materials. It could also be used as a benchmark for the development of speech processing algorithms: the original and the processed speech tokens could be fed as an input and their different outputs compared in order to see if the proposed technique has some improvement on predicted behavioural scores.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317521 (ICanHear). The authors would like to thank Dr. Raphael Koning for providing the base scripts from which this work was developed, Prof. Dr. Torsten Dau and Dr. Alexandre Chabot-Leclerc for their help in the calibration of the sEPSM ideal observer, and the anonymous reviewers of this manuscript for their helpful comments and feedback. Finally, our most sincere thanks to all the participants who volunteered for the study.

Endnotes

[1]Note that the ENV and TFS terminology is not strictly equivalent to that used in the study of Rosen (1992).

References

Allen, J. B. **(2005)**. "Articulation and intelligibility," Synthesis Lectures on Speech and Audio Processing **1**(1), 1–124.

ANSI **(1997)**. *American National Standard: Methods for Calculation of the Speech Intelligibility Index* Vol. 19 Acoustical Society of America New York, USA.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. and Wellekens, C. **(2007)**. "Automatic speech recognition and speech variability: A review," Speech Communication **49**(10), 763–786.

Bidelman, G. M. and Heinz, M. G. **(2011)**. "Auditory-nerve responses predict pitch attributes related to musical consonance-dissonance for normal and impaired hearing," J. Acoust. Soc. Am. **130**(3), 1488–1502.

Boersma, P. and Weenink, D. **(2014)**. "Praat: doing phonetics by computer,"'. Version 5.3.16. Date last viewed: 23-10-2014.
URL: *http://www.praat.org/*

Boothroyd, A. and Nittrouer, S. **(1988)**. "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**(1), 101–114.

Bradley, J. S. **(1986)**. "Predictors of speech intelligibility in rooms," J. Acoust. Soc. Am. **80**(3), 837–845.

Bruce, I. C., Léger, A. C., Moore, B. C. and Lorenzi, C. **(2013)**. "Physiological prediction of masking release for normal-hearing and hearing-impaired listeners," *in* 'Proceedings of Meetings on Acoustics' Vol. 19 Acoustical Society of America Montreal, Canada pp. 1–8.

Bruce, I. C., Sachs, M. B. and Young, E. D. **(2003)**. "An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses," J. Acoust. Soc. Am. **113**(1), 369–388.

Chi, T., Gao, Y., Guyton, M. C., Ru, P. and Shamma, S. **(1999)**. "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**(5), 2719–2732.

Cole, R., Yan, Y., Mak, B., Fanty, M. and Bailey, T. **(1996)**. "The contribution of consonants versus vowels to word recognition in fluent speech," *in* 'Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on' Vol. 2 IEEE Atlanta, Georgia, USA pp. 853–856.

Dau, T., Verhey, J. and Kohlrausch, A. **(1999)**. "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**(5), 2752–2760.

Drullman, R. **(1995)**. "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**(1), 585–592.

Elhilali, M., Chi, T. and Shamma, S. A. **(2003)**. "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun **41**(2), 331–348.

Ewert, S. D. and Dau, T. **(2000)**. "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**(3), 1181–1196.

Field, A., Miles, J. and Field, Z. **(2012)**. *Discovering Statistics Using R* SAGE California, USA.

Fogerty, D. and Kewley-Port, D. **(2007)**. "Investigating the consonant-vowel boundary: Perceptual contributions to sentence intelligibility," *in* 'Proceedings of Meetings on Acoustics' Vol. 2 Acoustical Society of America New Orleans, Louisiana, USA p. 060001.

Fogerty, D. and Kewley-Port, D. **(2009)**. "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**(2), 847–857.

Francart, T., Moonen, M. and Wouters, J. **(2009)**. "Automatic testing of speech recognition," Int. J. Audiol. **48**(2), 80–90.

Francart, T., Van Wieringen, A. and Wouters, J. **(2008)**. "Apex 3: a multi-

purpose test platform for auditory psychophysical experiments," J. Neurosci. Methods **172**(2), 283–293.

French, N. and Steinberg, J. **(1947)**. "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**(1), 90–119.

Ganong, W. F. **(1980)**. "Phonetic categorization in auditory word perception.," Journal of Experimental Psychology: Human Perception and Performance **6**(1), 110.

Gelfand, S. A. **(1998)**. "Optimizing the reliability of speech recognition scores," J. Speech Lang. Hear. Res. **41**(5), 1088–1102.

Green, D. M. and Birdsall, T. G. **(1964)**. *in* J. A Swets, ed., 'Signal Detection and Recognition by Human Observers' Wiley New York, USA.

Heinz, M. G. and Swaminathan, J. **(2009)**. "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," J. Assoc. Res. Otolaryngol. **10**(3), 407–423.

Hines, A. and Harte, N. **(2010)**. "Speech intelligibility from image processing," Speech Commun **52**(9), 736–752.

Hines, A. and Harte, N. **(2012)**. "Speech intelligibility prediction using a neurogram similarity index measure," Speech Commun **54**(2), 306–320.

Hornsby, B. W. **(2004)**. "The speech intelligibility index: What is it and what's it good for?," The Hearing Journal **57**(10), 10–17.

Hossain, M. E., Jassim, W. A. and Zilany, M. S. **(2016)**. "Reference-free assessment of speech intelligibility using bispectrum of an auditory neurogram," PloS one **11**(3), e0150415.

Jenkins, J. J., Strange, W. and Miranda, S. **(1994)**. "Vowel identification in mixed-speaker silent-center syllables," J. Acoust. Soc. Am. **95**(2), 1030–1043.

Jennings, S. G., Heinz, M. G. and Strickland, E. A. **(2011)**. "Evaluating adaptation and olivocochlear efferent feedback as potential explanations of psychophysical overshoot," J. Assoc. Res. Otolaryngol. **12**(3), 345–360.

Jennings, S. G. and Strickland, E. A. **(2012)**. "Evaluating the effects of olivocochlear feedback on psychophysical measures of frequency selectivity," J. Acoust. Soc. Am. **132**(4), 2483–2496.

Jørgensen, S. and Dau, T. **(2011)**. "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**(3), 1475–1487.

Jørgensen, S., Ewert, S. D. and Dau, T. **(2013)**. "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**(1), 436–446.

Kryter, K. D. **(1962)**. "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**(11), 1689–1697.

Lee, J. H. and Kewley-Port, D. **(2009)**. "Intelligibility of interrupted sentences at subsegmental levels in young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **125**(2), 1153–1163.

Lyon, R. and Shamma, S. **(1996)**. *in* 'Auditory computation' Springer New York, NY, USA pp. 221–270.

Mamun, N., Jassim, W. A. and Zilany, M. S. **(2015)**. "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (nopm)," IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(4), 760–773.

Miller, N. **(2013)**. "Measuring up to speech intelligibility," International Journal of Language & Communication Disorders **48**(6), 601–612.

Müsch, H. and Buus, S. **(2001)**. "Using statistical decision theory to predict speech intelligibility. i. model structure," J. Acoust. Soc. Am. **109**(6), 2896–2909.

Pavlovic, C. V. **(1987)**. "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**(2), 413–422.

Rosen, S. **(1992)**. "Temporal information in speech: acoustic, auditory and lin-

guistic aspects," Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **336**(1278), 367–373.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M. **(1995)**. "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.

Sidwell, A. and Summerfield, Q. **(1986)**. "The auditory representation of symmetrical cvc syllables," Speech Commun **5**(3), 283–297.

Smith, Z. M., Delgutte, B. and Oxenham, A. J. **(2002)**. "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**(6876), 87–90.

Steeneken, H. J. and Houtgast, T. **(1980)**. "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**(1), 318–326.

Stilp, C. E. and Kluender, K. R. **(2010)**. "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proceedings of the National Academy of Sciences **107**(27), 12387–12392.

Strange, W. and Bohn, O.-S. **(1998)**. "Dynamic specification of coarticulated german vowels: Perceptual and acoustical studies," J. Acoust. Soc. Am. **104**(1), 488–504.

Swaminathan, J. and Heinz, M. G. **(2012)**. "Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise," J. Neurosci. **32**(5), 1747–1756.

Wang, K. and Shamma, S. **(1995)**. "Spectral shape analysis in the central auditory system," Speech and Audio Processing, IEEE Transactions on **3**(5), 382–395.

Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P. **(2004)**. "Image quality assessment: from error visibility to structural similarity," Image Processing, IEEE Transactions on **13**(4), 600–612.

Williams, E. J. and Williams, E. **(1959)**. *Regression analysis* Vol. 14 Wiley New York, NY, USA.

Young, E. D. **(2008)**. "Neural representation of spectral and temporal information in speech," Philosophical Transactions of the Royal Society B: Biological Sciences **363**(1493), 923–945.

Zhang, X., Heinz, M. G., Bruce, I. C. and Carney, L. H. **(2001)**. "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," J. Acoust. Soc. Am. **109**(2), 648–670.

Zilany, M. S. and Bruce, I. C. **(2006)**. "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," J. Acoust. Soc. Am. **120**(3), 1446–1466.

Zilany, M. S. and Bruce, I. C. **(2007)**. "Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery," *in* 'Neural Engineering,

2007. CNE'07. 3rd International IEEE/EMBS Conference on' IEEE Kohala Coast, HI, USA pp. 481–485.

Zilany, M. S., Bruce, I. C. and Carney, L. H. **(2014)**. "Updated parameters and expanded simulation options for a model of the auditory periphery," J. Acoust. Soc. Am. **135**(1), 283–286.

Zilany, M. S., Bruce, I. C., Nelson, P. C. and Carney, L. H. **(2009)**. "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics," J. Acoust. Soc. Am. **126**(5), 2390–2412.

Zilany, M. S. and Carney, L. H. **(2010)**. "Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics," J. Neurosci. **30**(31), 10380–10390.

Figure Captions

Figure 1. Overview of the materials and methods used in this study. Dashed arrows represent the reference (clean) signal. Dotted arrows represent the noise signal. Dashed-and-dotted arrows represent the degraded (clean + noise) signal.

Figure 2. Example segmentation of the word *bot*. The time signal is shown in the top part. The spectrogram is shown in the bottom part.

Figure 3 (color online). Example ENV (left column) and TFS (right column) neurograms for the word *bot*. Top row corresponds to the reference condition (in quiet). Middle row corresponds to the condition SNR = 0 dB. Bottom row corresponds to the condition SNR = -12 dB. Notice how information is represented differently by the ENV and TFS neurograms.

Figure 4 (color online). Example ES neurograms (left column) and cortical representation (right column) for the word *bot*. Top row corresponds to the reference condition (in quiet). Middle row corresponds to the condition SNR = 0 dB. Bottom row corresponds to the condition SNR = -12 dB. Cortical representations correspond to an example temporal modulation rate of 5.65 Hz and a spectral modulation filter scale of 1 cycles/oct.

Figure 5 (color online). Different objective measures for 15 randomly picked (sample) words. It is important to mention that model parameters were not individualized and that although all objective measure have the same range (0 to 1), they are of a different nature.

Figure 6. Distribution of different objective measures of the complete set averaged across words. Circles represent the mean. The dotted line represents the fitted straight line. Crosses represent outliers. Note that although all objective measure have the same range (0 to 1), they are of a different nature.

Figure 7 (color online). Scatter plots of phoneme scores vs different objective measures averaged across participants at a word level (thus each point is a word). Reported correlations and F-ratios were calculated with $\alpha = 0.05$ and significant.

Figure 8 (color online). Correlation between the behavioural scores and the computed metrics (per phoneme). Reported correlations were calculated with $\alpha = 0.05$.